Obtaining prominence judgments from naïve listeners

Denis Arnold [1], *Petra Wagner* [2], *Bernd Möbius* [3]

[1] Quantitative Linguistics, University of Tübingen, Germany
[2]Faculty of Linguistics and Literature, University of Bielefeld, Germany
[3] Department of Computational Linguistics and Phonetics, Saarland University, Germany
denis.arnold@uni-tuebingen.de, petra.wagner@uni-bielefeld.de, moebius@coli.uni-saarland.de

Abstract

In this paper we examine different approaches to obtain judgments of perceptual prominence. It discusses the use of different scales, the influence of the linguistic level on which the prominence is rated and the normalisation of prominence judgments. We propose the use of a multilevel scale to obtain prominence judgments. It seems that naïve listeners can rate word prominence better than syllable prominence, resulting in better correlations to acoustics. It is shown that normalization should be applied to the obtained ratings.

Index Terms: prosody, prominence, methods, normalisation, acoustic correlates

1. Introduction

1.1. Motivation for prominence research

It is widely accepted, that prominence is a perceptual construct, that describes the perceived strength of a given linguistic unit to its neighbours. The question remains, wether prominence is gradual or categorical, wether prominence should be labeled on word or syllable level and wether there is something like an "absolute" prominence or if prominence has to be seen in its context. Unfortunately a lot of papers dealing with prominence do not give or refer to a definition of prominence. As it has been pointed out by [1] a lot of the results found by the different studies on prominence might be influenced by the approach the authors chose. For our studies we refer to the definition given in [2].

1.2. Different approaches to capture prominence

A lot different methodological approaches have been used in prominence research. Experiments on natural [3, 4] and manipulated stimuli [5, 6] were conducted. Production experiments were carried out [1] as well as work using corpora with annotated prominence [7, 8]. Some studies dealt with the prediction of prominence [9] and some focused on the automatic annotation of prominence from the signal [10, 11] and others on the systematic differences between automatic and human annotated prominence [8]. Also studies examining the influence of linguistic knowledge on the perceptual prominence were conducted [4, 12, 13]. All the mentioned studies use a lot of different scales, measure prominence on different linguistic levels and also have different concepts on prominence. Additionally the research is conducted with a lot of different languages. While different approaches, like experiment vs. corpus work, are valuable to examine different questions, one has to be very careful when comparing the results of the studies that use different methods to capture prominence.

2. Evaluations of different scale to obtain prominence

A lot different rating scales have been employed in prominence research. A lot studies use a binary scale e.g. [14], [15] and [16]. The clear advantage of this procedure is, that it is easy to use for the raters. It is argued that with *n* raters one gets a n-level scale of prominence. However, this leads to another problem. If one uses the number of raters that say a given unit is prominent, one confuses the amount of prominence with the confidence of the rating. With a multilevel scale one can see, how much prominence is assigned to a given unit and how confident the raters are. Our data [17] and [18] indicate, that high confidence is not equal to extreme prominence ratings and vice visa. A lot of different multilevel scale were used in the literature including a 3-point [8], 4-point [19], 11-point [20] and 30-point scale [3, 12, 13]. Two studies that focused on the use of scales found contradicting results [21], [19]. Grover et al. found that scales with more levels result in more reliable results [21] , while Jensen and Tondering prefer the use of a 4-point scale [19]. While the results found with the three tested scales - binary, 4-point and 31-point scale - do not differ much, the authors say, that the 31-point scale is harder to use for naïve listeners, that the range of that scale can not be utilized by naïve listeners and that one finds less extreme results with the 31-point scale. In [17] we presented data that supports the use of multilevel scales like a 11-point or 31-point scale over the use of a 4-point scale and a continuos scale for the rating of prominence. We did not find that scales

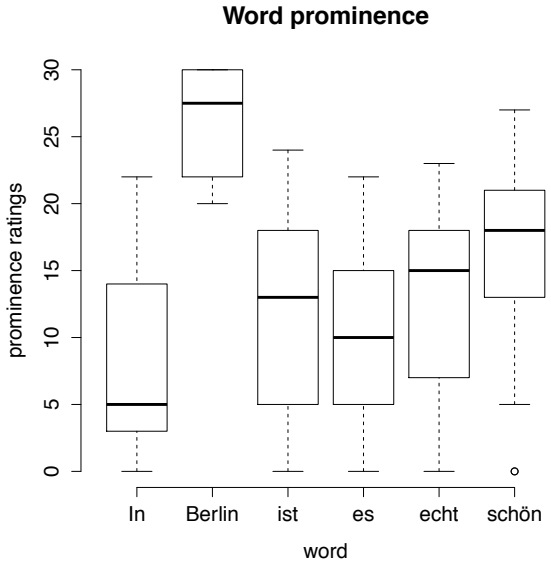

Figure 1: *Word prominence rating of a sentence in study [18]. In Berlin ist es echt schön - It is really nice in Berlin.*

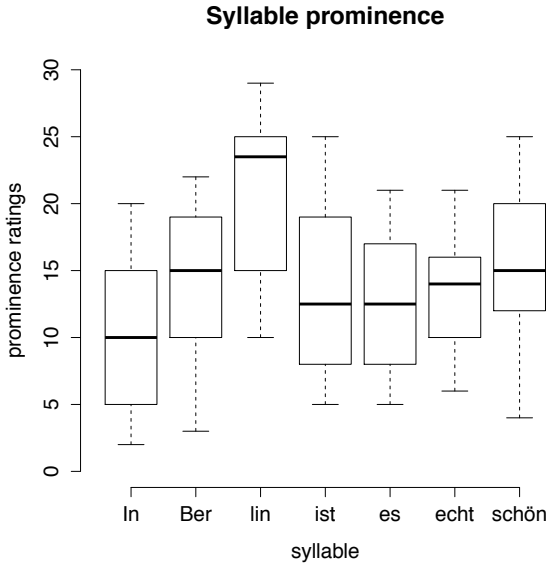

Figure 2: *Syllable prominence of a sentence in study [18]. In Berlin ist es echt schön - It is really nice in Berlin.*

with many levels can not be utilized by our subjects, that more scale levels results in less extreme results and that many scale levels are harder to use for naïve listeners.

3. Prominence on different linguistic levels

The different aspects of perceptual prominence have been examined in different studies on word level and syllable level. A systematic variation of the linguistic unit was only reported in a small experiment found in [14]. In [18] we presented data which showed that there is no simple relation beween the prominence rated on word level and prominence ratings. We found that complex interactions influence the assigned prominence on the different linguistic levels. Figures 1 and 2 show the prominence of the same sentence, once rated by a group on word level and once rated by another group on syllable level. The word prominence of "Berlin" is much higher than the syllable prominence of "lin", which carries the word accent. There is an great difference in the first word "In" which is mostly influenced by the context. We found these two effects in most of the data. One can say that the prominence is often greater as the prominence of the syllable that carries the word accent. There are often units that show great differences in prominence because of the changes in their direct neighbours. Table 1 shows that the correlations between acoustic features and the prominence ratings are greater for the ratings on word level. Combined with the finding of lower costs (c.f. [18]) for the rating on word level, one can conclude, that rating on word level is easier for naïve listeners.

4. Normalisation

The normalisation of prominence ratings is rarely reported. Two examples are [4, 24]. In [23] different approaches of normalization of prominence ratings are evaluated on one data set. The study found that normalization of prominence ratings improves the correlation between prominence ratings and acoustics significantly. We applied a normalisation to the data of [17] and [18]. Therefore we used a z-transformation, which is described by equation 1.

$$Z_n = \frac{R_n - \mu}{\sigma} \tag{1}$$

R_n is the prominence at the syllable n. μ is the mean and σ the standard deviation of all ratings.
All statistic calculations were carried out by means of the free statistics program and language R [22].

Table 1: *Acoustic correlates for the raw data of study [18]*

	Word prominence	Syllable prominence
Dauer	.69	.41
	$p < .001$	$p < .001$
Maximum f0	.54	.40
	$p < .001$	$p < .001$
Intensitt	.53	.39
	$p < .001$	$p < .001$

4.1. Results

We found several advantages of the normalisation of the prominence ratings. The correlations between acoustics and prominence ratings improve in most cases in both data sets. Table 3 shows the correlations between acoustics and the normalised prominence ratings from [18]. In [17] the priming effect could only be replicated using the 31-point scale (c.f. Table 3). After the normalisation the effect was replicated with the 31-point and the 11-point scale. The effect is not significant for the ratings obtained with the 4-point and continuos scale. Figure 3 shows the raw prominence ratings of a sentence from the data of [18]. The last syllable carries the word accent of the last word. One would expect that this syllable would receive a higher prominence rating the first syllable of the same word. Table 4 shows the normalised prominence ratings for the same sentence. It shows, that the ratings are much more in line with the linguistic expectations after the normalization.

4.2. Discussion

The general findings from [17] and [18] are not changed by the normalization. The normalization shows some positive effects. The correlations between the acoustic features and the normalized prominence ratings are better after the normalization for both data sets. This is in line with the findings from [23]. As expected normalization compensates artefacts in the ratings, as shown in figures 3 and 4. Since the effect of priming did not vanish, these results give further support that priming of prominence pattern works. After the normalization the ratings obtained with the 11-point scale also showed a priming effect.

5. Conclusions

We conclude, that the use of a scale with more levels enables interesting insights into the perception of prominence by naïve listeners. In difference to ratings with a binary scale, one can observe the prominence and the confidence. It shows that the subjects show a higher agreement on the prominence of certain units. These do not have to be the most prominent units in the sentence.

Table 2: *Results of the priming in [17] using the raw data and normalised data with the four different scales.*

	4-point	11-point	31-point	continuous
Raw	W = 140.5	W = 185.5	W = 229	W = 143.5
	p = .49	p = .46	$p < .05$	p = .56
Z	W = 175	W = 342	W = 229	W = 171
	p = .34	$p < .05$	$p < .05$	p = .39

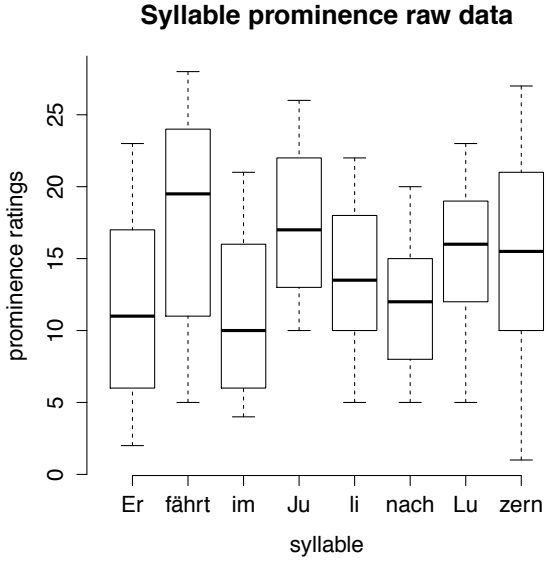

Figure 3: *Raw syllable prominence ratings of a sentence in study [18]. Er fährt im Juli nach Luzern - He will go to Luzern in July.*

It will be interesting to investigate these variations more in detail. The ratings obtained with multilevel scales show good correlations to the acoustics and a good detection of rating differences. We did not find any disadvantages with respect to the difficulty for the raters. The priming effect from [13] was only replicated using the 11-point and 31-point scale, while the replication failed with the 4-point and continuos scale.
Word prominence is easier to rate than syllable prominence for naïve listeners. Since the results of prominence obtained on word and syllable level differ significantly, one should be careful when comparing results from studies using different levels as a reference.
We find that a normalisation of the prominence ratings shows several advantages. The acoustic correlates get stronger, as well as the discrimination in rating differences. It shows that normalization results in less artifacts in the prominence ratings. The effect that the subjects

Table 3: *Acoustic correlates for the normalised data of study [18]*

	Word prominence	Syllable prominence
Duration	.70	.39
	$p < .001$	$p < .001$
Maximum f0	.54	.43
	$p < .001$	$p < .001$
Intensity	.54	.44
	$p < .001$	$p < .001$

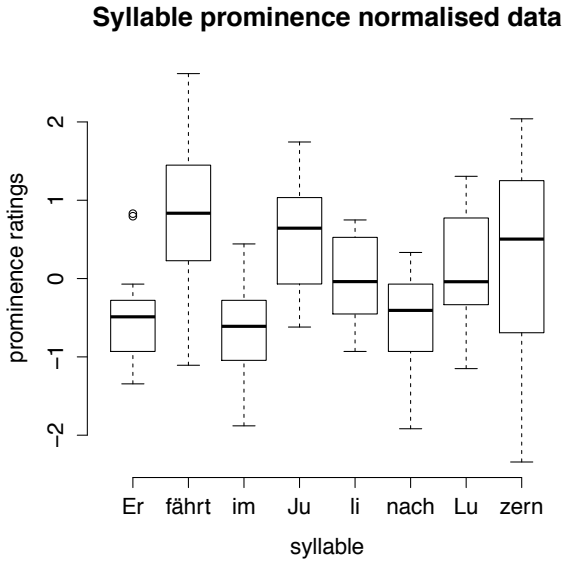

Figure 4: *Normalised syllable prominence ratings of a sentence in study [18]. Er fährt im Juli nach Luzern - He will go to Luzern in July..*

agree much more on certain units does not disappear with the normalization. The normalization does not change the general findings of [17, 18].

6. References

[1] Watson, D. G., Arnold, J. E. and Tanenhaus, M. K., "Tic Tac TOE: Eects of predictability and importance on acoustic prominence in language production", Cognition 106, 15481557, 2008.

[2] Wagner, P., "Wahrnehmung und Vorhersage deutscher Betonungsmuster", Universität Bonn. PhD-Thesis, 2002. Online: http://hss.ulb.uni-bonn.de/2002/0054/0054.htm, accessed on 29 Mar 2011.

[3] Fant, G und Kruckenberg, A., "Preliminaries tothe study of Swedish prose reading and reading style", STR-QPSR, 2/1989 KTH, Stockholm, 1-83, 1989.

[4] Eriksson, A., Grabe E. und Traunmüller, H., "Perception of syllable prominence by listeners with and without competence in the tested language". Proceedings Speech Prosody 2002, Aix-en-Provence, 275-278, 2002.

[5] Gussenhoven, C., Repp, B. H., Rietfeld, A. and Terken, J., "The perceptual prominence of fundamental frequency peaks", Journal of the Acoustical Society of America 102, 30093022, 1997.

[6] Gussenhoven, C. and Rietveld, T., "On the speaker-dependence of the perceived prominence of F0 peaks," Journal of Phonetics 26, 371380, 1998.

[7] Kochanski, G., Grabe, E., Coleman, J., and Rosner, B., "Loudness predicts prominence: fundamental frequency lends little". Journal of the Acoustical Society of America 118, 1038-1054, 2005.

[8] Goldman, J.-P., Auchlin, A., Roekhaut, S., Simon, A. C., and Avanzi M., "Prominence perception and accent detection in French. A corpus-based account", Proceedings of Speech Prosody 2010, Chicago, 2010.

[9] Widera, C., Portele, T. and Wolters, M., "Prediction of word prominence", Proceedings of Eurospeech 1997, 999-102, 1997.

[10] Tamburini, F. "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system", Proceedings of Eurospeech 2003, 129-132, 2003.

[11] Wang, D. and Narayanan, S.; , "An Acoustic Measure for Word Prominence in Spontaneous Speech." IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.2, 690-701, 2007.

[12] Wagner, P., "Great Expectations - Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates", Proceedings of Interspeech 2005, Lisbon, 2381-2384, 2005.

[13] Arnold, D., Wagner, P. and Möbius, B., "The effect of priming on the correlations between prominence ratings and acoustic features", Proceedings of Speech Prosody 2010, Chicago, 2010.

[14] Streefkerk, B., "Prominence - Acoustic and lexical/syntactic correlates", Utrecht: LOT, 2002.

[15] Cole, J., Mo, Y. and Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence", Laboratory Phonology 2010, 1:2, 425-452, 2010.

[16] Mo, Y., Cole, J. and Lee, E.-K. "Naïve listeners prominence and boundary perception", Proceedings of Speech Prosody 2008, Campinas, 2008.

[17] Arnold, D., Wagner, P., and Möbius, B. "Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners", Proceedings of ICPhS 2011, Hong Kong, 2011.

[18] Arnold, D., Möbius, B., and Wagner, P., "Comparing word and syllable prominence rated by naïve listeners", Proceedings of INTERSPEECH 2011, Florence, 2012.

[19] Jensen, C. and Tøndering, J. , "Choosing a Scale for Measuring Perceived Prominence", Proceedings of Interspeech 2005, Lisbon, 2385-2388, 2005.

[20] Turk, A. E. and Sawusch, J. R., "The processing of duration and intensity cues to prominence", Journal of the Acoustical Society of America 99, 3782-3790, 1996.

[21] Grover, C., Heuft, B. und van Coile, B., "The reliability of labeling word prominence and prosodic boundary strength", Proceedings of the ESCA Workshop on Intonation, Athens, 165-168, 1997.

[22] R Development Core Team. "R: A language and environment for statistical computing.", R Foundation for Statistical Computing, Vienna, 2012.

[23] Sappok, C. and Arnold, D., "On the Normalization of Syllable Prominence Ratings", Proceedings Speech Prosody 2012, Shanghai, 2012.

[24] Liljencrants, J., "Judges of prominence", Fonetik 99: Proceedings from the Twelfth Swedish Phonetics Conference, Göteborg, 101-107, 1999.